Ahrii Kim[1]

[1]Independent Researcher

August 11, 2025

Multi-agentMT: Deploying AI Agent in the WMT25 Shared Task

Ahrii Kim
Independent Researcher / Seoul, Korea
ahriikim@gmail.com

Abstract

We introduce our model, referred to as Multi-agentMT, for participation in the WMT 25 General Machine Translation Shared Task. This model operationalizes the notion of an AI Agent by employing a multi-agent workflow known as Prompt Chaining (Briva-Iglesias, 2025) alongside the automatic MQM (Multidimensional Quality Metrics) error annotation framework designated as RUBRIC-MQM (Kim, 2025). Our primary submission is developed through the **Translate-Postedit-Proofread** paradigm, whereby the positions of the errors are clearly marked and enhanced throughout the process. Our study suggests that a semi-autonomous agent scheme in Machine Translation is viable with an older and smaller model in some language pairs, resulting in comparable results with 2.3x faster speed and only 2% of the budget.

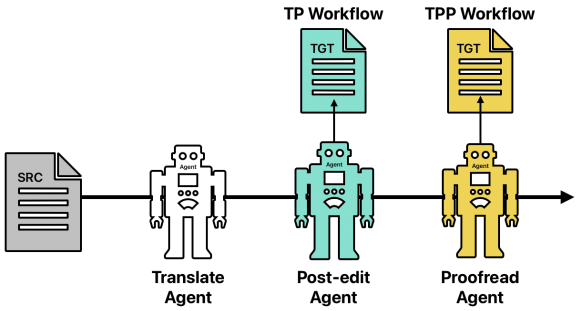


Figure 1: The Prompt Chaining architecture of Multi-agentMTof Translate-Postedit-Proofread. Our submission comprises two methodologies: the Translate-Postedit, referred to as the TP Workflow, and the Translate-Postedit-Proofread, designated as the TPP Workflow. The TPP Workflow is presented as our primary submission. Within this process, each agent sequentially receives the output from its predecessor as input, thereby iteratively enhancing the translation quality.

1 Introduction

An AI Agent is a computational system that operates autonomously, guided by environmental observations, and frequently incorporates adaptive learning abilities (Russell and Norvig, 2010). Large Language Models (LLM), essential in advancing AI Agents, boast enhanced reasoning, contextual comprehension, and adaptable workflows across diverse tasks, notably including Machine Translation (MT) (Briva-Iglesias, 2025). In this context, Briva-Iglesias (2025) investigated a multi-agent MT system with four agents: Translator, Fluency Reviewer, Adequacy Reviewer, and Editor, showing its promising future potential. Motivated by this method, we are participating in this year's WMT (The Conference on Machine Translation) General Task utilizing an AI multi-agent workflow. **Our goal is to create a smaller model that outperforms larger ones, thus demonstrating the potential of AI Agent in MT with significantly reduced costs.**

The main aim of the competition this year is to translate texts from a wide variety of languages, fields, genres, and formats. We have taken part in a **multilingual subtask** covering 30 languages, with Czech, English, and Japanese as source languages. Adhering to prompt engineering without prioritizing specific languages as primary, our model can be categorized as both a **contrastive** and **unconstrained** type.

A significant issue was the adherence to the prior year's convention, wherein the dataset comprised document boundaries, resulting in segments characterized by a series of sentences and paragraphs, delineated by one or two newline characters. This structure yielded a total of 29,957 segments or 102,060 paragraphs. Our initial submission, which implemented translation at the segment level, failed to respect these boundaries, often either merging them or omitting some paragraphs entirely, particularly in the TPP Workflow (this issue is elaborated in § 5.2). Consequently, for such instances,

we partitioned the segments into individual paragraphs and processed them independently in the inference process. Apart from this specific instance, the majority of the translations are executed on a per-segment basis.

As per the official results reported by WMT **our model shows expertise in certain languages, specifically Icelandic (`is_IS`) and Serbian (`sr_Cyrl_RS`), and Swahili (`mas_KE`)**, the underlying factors contributing to this success have not yet been identified. It is particularly meaningful considering that our baseline model is not the most updated version. Elucidating the mechanisms through which our model attains high-quality translations across these unrelated languages would contribute significantly to advancing research in the domain of multilingualism of LLM.

The structure of the paper is arranged as follows: Section § 2 provides a detailed architecture of multi-agents, whose performance is measured in Section § 3 utilizing the WMT24++ dataset (Deutsch et al., 2025), and Section § 4 discusses the results. Appendix provides a detailed explanation of the prompt lines.

2 System Overview

2.1 Design

AI Agents facilitate dynamic workflows through configurable architectures. We utilize the concept of Prompt Chaining, which forms an orderly and consecutive process where each step's result feeds into the next, promoting systematic reasoning and ongoing refinement (Briva-Iglesias, 2025). While theoretically, the iterative refinement could enhance translation quality, cost considerations have led us to adopt a unidirectional configuration. Consequently, we examine two types of multi-agent workflows: Translate-Postedit (**TP Workflow**) and Translate-Postedit-Proofread (**TPP Workflow**) as depicted in Figure 1. Each scenario is submitted for the competition.

2.2 Translate Agent

The model is tasked with translating the provided source text according to the officially provided prompt from the organizer. For reasons of cost-effectiveness, it would be feasible to obtain translations from more economical platforms such as Google Translate or DeepL. Nonetheless, *our preliminary research indicates that a high-quality initial translation produces superior outcomes*.

Algorithm 1: post_edit_translation(response, tgt_text)

```
Input: response, tgt_text
Output: corrected
raw ← response["content"] or ""
corrected ← tgt_text
MIN_SAFE_SPAN_LEN ← 2
try:
    safe_response ← sanitize_response(raw)
    parsed ← JSON parse of safe_response
    if parsed is a dictionary then
        forall span in parsed do
            info ← parsed[span]
            if info is not a dictionary then
                continue
            suggestion ← clean_suggestion(
                info["suggestion"].strip() )
            if span.lower() == "no-error"
                or suggestion is empty
                or suggestion == span then
                continue
            if length(span) < MIN_SAFE_SPAN_LEN then
                continue
            space ← " "
            pattern_space ← space + escape(span) + space
            (corrected, count) ← regex_subn(pattern_space,
                space + suggestion + space, corrected)
            if count == 0 then
                pattern_general ← escape(span)
                (corrected, _) ← regex_subn(pattern_general,
                    suggestion, corrected)
except:
    corrected ← tgt_text
corrected ← preserve_paragraph(tgt_text, corrected)
return corrected
```

2.3 Post-edit Agent

The model is tasked with post-editing the provided translation with reference to the source text. RUBRIC-MQM (Kim, 2025) serves as the foundational framework, which is a LLM-as-judge that discerns MQM-style (Multidimensional Quality Metrics) error categories, severities, and error spans akin to its counterpart GEMBA-MQM (Kocmi and Federmann, 2023). In elucidating RUBRIC-MQM, it has demonstrated robustness in determining error categories with heightened precision, particularly in conjunction with MAJOR and MISTRANSLATION labels, and offers improved differentiation between flawless and flawed sentences (Kim, 2025).

Four key aspects are revised from the original version. Firstly, in order to transform the LLM-as-judge into a post-editor, we instruct the model to propose better translations for each identified error span. Secondly, the decision is made to reduce the severity scale from 100 to 4. This adjustment is made because the severity level is not our main focus, but Kim (2025) highlights the rubric scheme's importance for improving model performance. As a result, the comprehensive rubric is streamlined within the prompt. Thirdly, while maintaining the original in-context learning examples, we modify one instance from an English-to-German translation direction to a Japanese-to-Korean one, thereby facilitating a multilingual configuration for an X-Y direction. Finally, the model is not afforded the option of NO-ERROR, ensuring that it identifies at least one error. This methodology is deemed necessary within our workflow, as selecting NO-ERROR by the model would preclude any enhancements to the translation. Consequently, STYLE is introduced as a new category label, thereby ensuring the model consistently offers corrections.

As a postprocessing step, the suggested translations are manually interchanged as shown in Algorithm 1. For each error span and its suggestion, the algorithm attempts a two-step substitution:

- **Space-sensitive substitution**: It first tries to replace the span only if it appears surrounded by spaces (e.g., replacing "␣mistake␣" with "␣correction␣"). This avoids incorrect substitutions in partial words.
- **Fallback substitution**: If no replacement occurs in step 1 (i.e., count == 0), it performs a more general substitution, replacing the span wherever it appears.

This allows the algorithm to fall back to broader replacements only when necessary, preserving accuracy while ensuring corrections are applied. The revised sentence is deemed the ultimate translation result of TP Workflow.

2.4 Proofread Agent

The model is involved in the examination and refinement of the provided translation, and we employ a Chain-of-Thought (CoT) prompting strategy (Wei et al., 2022). Initially, the model is guided to identify errors, followed by a request to propose five alternative phrases that prioritize fluency of its own and alignment with the source text. Ultimately, the most suitable translation is selected by the model as the final response. It is expected that this methodology will improve any awkward expressions resulting from manual alterations in the preliminary phase. The final output is deemed the ultimate translation result of TPP Workflow.

2.5 Model Architecture

For the baseline model, we utilize `GPT-4o-mini-2024-07-18`, a proprietary model developed by OpenAI (OpenAI et al., 2023), across all agents, which is configured with a temperature= 1 and a max_token= 1024. Although configuring the temperature to 1 is not deemed optimal for reproducibility, our series of iterative pilot studies has indicated that such a configuration facilitates the model's exploration of a wide spectrum of errors and modifications, thereby enhancing the outcomes. The model was executed between June 19 and July 3, 2025.

3 Experiment

This section outlines the relative effectiveness of our model as a small, older version within the languages discussed earlier: Icelandic, Serbian, and Swahili (Kenya). We utilize the WMT24++ dataset (Deutsch et al., 2025), which consists of an English source text and 55 corresponding translations in different languages, alongside up to two reference translations, specifically a human reference and a post-edited version. The source segments for each language pair total 960 after removing low-quality segments.

Our translations are produced from the TP and TPP Workflows, and scores of BLEU (Papineni et al., 2002), ChrF (Popović, 2015), TER (Snover et al., 2006), three of which are calculated from

Language	Metric	Translate	Postedit	Proofread
Icelandic	BLEU	18.33	17.91 (-0.42)	18.00 (+0.09)
	ChrF	43.42	42.96 (-0.46)	43.55 (+0.59)
	TER	67.49	69.61 (+2.12)	70.28 (+0.67)
	COMET	78.75	76.90 (-1.85)	79.22 (+2.32)
	COMET Kiwi	75.74	73.89 (-1.85)	76.41 (+2.52)
Serbian	BLEU	23.12	21.92 (-1.20)	20.39 (-1.53)
	ChrF	49.96	48.26 (-1.70)	46.02 (-2.24)
	TER	63.79	67.10 (+3.31)	69.73 (+2.63)
	COMET	82.49	79.31 (-3.18)	81.42 (+2.11)
	COMET Kiwi	80.66	77.73 (-2.93)	80.69 (+2.96)
Swahili (Kenya)	BLEU	20.60	20.10 (-0.50)	20.22 (+0.12)
	ChrF	49.99	49.60 (-0.39)	50.32 (+0.72)
	TER	67.94	71.50 (+3.56)	71.23 (-0.27)
	COMET	78.16	77.02 (-1.14)	78.39 (+1.37)
	COMET Kiwi	77.15	75.58 (-1.57)	77.40 (+1.82)

Table 1: The individual performance scores of the Multi-agentMTsystem. "Translate" is done by Gemini-1.5-Flash. Colored values indicate the score change from the previous stage: positive or negative.

SacreBLEU (Post, 2018), COMET (Rei et al., 2020), and COMETKiwi (Rei et al., 2022) are calculated based on two references. Additionally, to adequately quantify the extent of modifications by each process, TER scores are computed between the workflows. For the purpose of evaluating cost-efficiency, the time and number of tokens expended are computed for each scenario.

For this experiment, several configuration adjustments were implemented. Initially, the Translate Agent was replaced with the pre-existing translation of Gemini-1.5-Flash for efficiency. Furthermore, the temperature was set to 0 to ensure the reproducibility of the results.

For an extended analysis, we evaluate the larger, latest baseline model utilizing GPT-o3-pro, assessing both its performance and cost-effectiveness. In this context, En-Is translation is exclusively employed due to budget constraints.

4 Result

4.1 Performance

As demonstrated in Table 1, the metric scores typically decrease following post-editing and subsequently increase after proofreading. A primary inquiry is whether these alterations are beneficial or detrimental. Superficially, continual modifications are evidenced by the consistently ascending TER scores, indicating a positive outcome whereby the prompting instruction is effective in facilitating ongoing changes rather than yielding a "no-error" response. Furthermore, from a semantic perspective, the adjustments indicated by the two COMET-type scores remain within a 2.0-point range, thereby suggesting that these modifications are not negative, at the very least.

To gain a deeper understanding of the observed trend, we have calculated the extent of edits conducted between two agents with TER scores. As indicated in Table 2, the findings reveal that the most significant edits occur between the Postedit-Proofread agents (PE-PR), approximately 2.25× more than the TP process (Trans-PE). Furthermore, when comparing the interactions between Translate and Proofread agents (Trans-PR), it is observed that, on average, 33.3% of the edits are executed. This result suggests that **the final translation within the TPP Workflow diverges considerably from both the original and the post-edited sentences.**

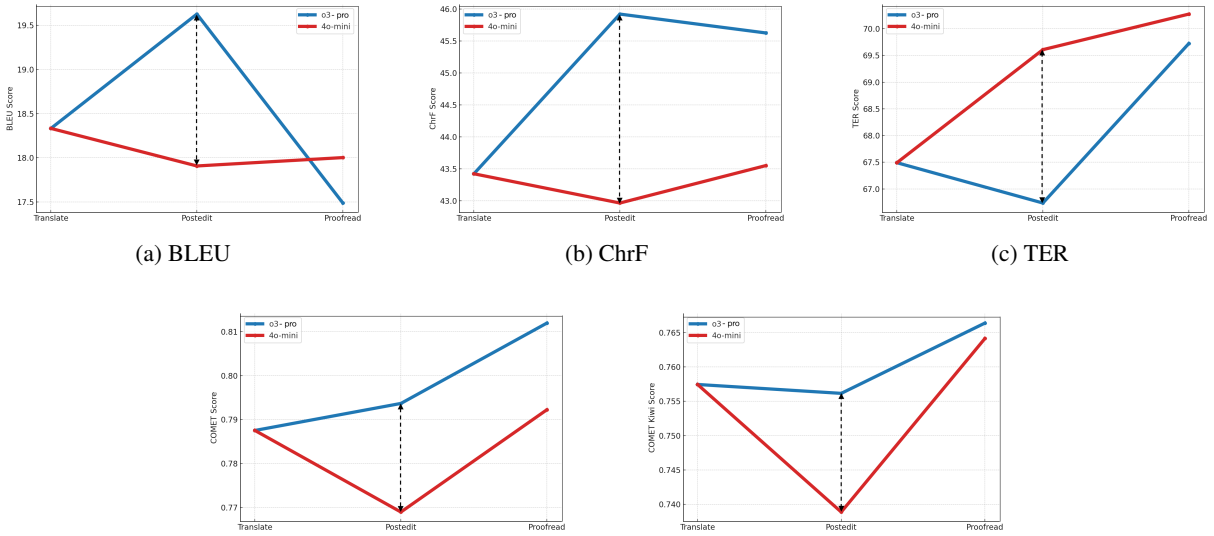
Language	Trans-PE	PE-PR	Trans-PR
Icelandic	13.12	29.58	31.89
Serbian	13.69	31.28	33.43
Swahili	16.42	31.20	34.68

Table 2: The edit distance is measured between each stage in our translation workflow. Specifically, 'Trans', 'PE', and 'PR' refer to each agent responsible for Translation, Post-edit, and Proofread, respectively. Although not incorporated in our official submission, we also compute the distance between Translate vs. Proofread.

By synthesizing the aforementioned findings, it can be inferred that **the model primarily makes alterations at the phrase level while maintaining the original structure**. The determination of whether such alterations are beneficial or detrimental necessitates human assessment. Furthermore,

Method	Translate Tokens	Post-edit Tokens	Proofread Tokens	Cost ($)
Translate	10,728,181	–	–	2.00
Post-edit	–	58,548,177	–	10.89
Proofread	–	–	15,874,680	2.95
TP (Translate + Post-edit)	10,728,181	58,548,177	–	12.89
TPP (Translate + Post-edit + Proofread)	10,728,181	58,548,177	15,874,680	15.84
GPT-4o	10,728,181⋆	–	–	33.26
GPT-o3-pro	10,728,181⋆	–	–	806.48

Table 3: Comparison of token usage and cost for our two workflows (TP, TPP) and bigger GPT models. Assuming that the larger model utilizes a comparable number of tokens as outlined in ⋆, our models are projected to consume a greater quantity of tokens while incurring lower costs, thereby achieving comparable translation quality.



(a) BLEU (b) ChrF (c) TER

(d) COMET (e) COMET Kiwi

Figure 2: Performance of o3-pro vs. 4o-mini measured by five automatic metrics for the English-Icelandic translation. A dotted line delineates the discrepancy observed in the Postedit Agent between the two models.

in the context of language pairs, the model demonstrates superior performance in Serbian, underscoring its enhanced efficacy in languages with higher resources.

4.2 Cost-efficiency

This section presents an analysis of the cost-efficiency of Multi-agentMTin the context of the official task. Table 3 provides a comprehensive overview of the quantity of tokens utilized, while Table 4 delineates the pricing structure for each respective model. Our model has exhibited an average input-output ratio of $0.08 : 0.92$, upon which the cost calculation is predicated. The findings reveal that within our framework, a considerable proportion of tokens is designated for the Postedit Agent, accounting for 68.75% of total usage, with an expenditure of $10.89 out of $15.84. If we postulate that the efficacy of our model is analogous to the straightforward translation processes of larger models (and assuming that the equivalent number of tokens is expended), the data indicates that **identical quality can be attained in select languages with merely 2% of their spending**.

	4o-mini	4o	o3-pro
Input	$0.15	$2.50	$20.00
Output	$0.60	$10.00	$80.00

Table 4: Pricing details per 1M input/output tokens for various GPT models (OpenAI).

5 Further Study

5.1 4o-mini vs. o3-pro

Our model is evaluated in comparison to the more extensive and recent GPT-o3-pro model (`o3-pro-2025-06-10`), specifically in terms

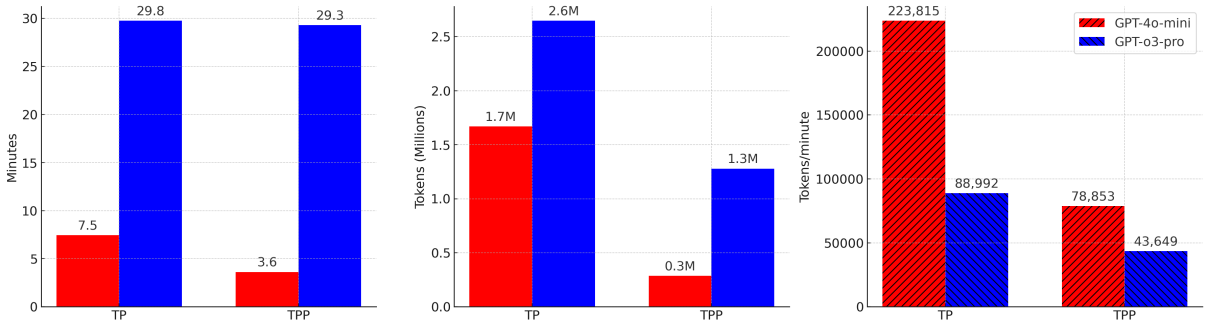


Figure 3: Token consumption per 1 million (1M) and time in minutes for 4o-mini and o3-pro models during each agent stage. The final figure computes throughput of the models in tokens per minute. 4o-mini is 2.3 times faster than o3-pro in our framework.

of performance and cost-effectiveness for the English-Icelandic language pair. Figure 2 illustrates the scores derived from automatic metrics. Although the ultimate scores within the TPP Workflow vary depending on the metrics used, it was observed that **the Postedit agent demonstrates increased activity in morphological and semantic aspects when engaging with the o3-pro model as opposed to the 4o-mini model**.

Figure 3 depicts the number of tokens utilized, the time taken, and consequently the throughput of the models. GPT-4o-mini demonstrated a substantial performance advantage over GPT-o3-pro in both the TP and TPP processes. In the TP process, 4o-mini exhibited a speed approximately 4 times greater and utilized only 63% of the tokens compared to o3-pro. Additionally, its processing velocity exceeded 2.5 times the number of tokens per minute. A similar pattern was observed in the TPP process, where 4o-mini accomplished the task approximately 8 times faster while reducing token usage to 22%, effectively doubling the token processing speed. **These findings underscore the superior computational efficacy of GPT-4o-mini for translation applications where time and cost efficiency are of critical importance**.

5.2 Qualitative Study

This section examines practical instances of the Multi-agentMTframework to elucidate its operational dynamics. Due to limitations in space, examples are provided in the Appendix. The selected example presented in Table 5 underscores the scenario where the Post-edit Agent's output is suboptimal, yet the Proofread Agent successfully mitigates this error. **The key aspect of Multi-agentMTis the Post-edit Agent's ability to pinpoint revision areas, even if its annotations result in incorrect translations, unlike solely using larger models.** In the cited instance, the Post-edit Agent retained the term "*blast*" from the source, which the Proofread Agent subsequently amended by altering its surrounding context.

On the negative side, it has been observed that the Proofread Agent exhibits a tendency to hallucinate long sentences by omitting portions of the original text when the translation provided is lengthy, consequently disregarding the document boundaries inherent to the original source sentences. As in the sample translation presented in Table 6, approximately half of the sentences appear to be omitted in the Proofread Agent's output. This phenomenon occurs with considerable frequency, particularly in the context of longer sentences. We intend to examine this issue more thoroughly in future research.

6 Conclusion

Our investigation on WMT24++ (Deutsch et al., 2025) indicates that the Post-edit Agent is integral to the translation process, underscoring the significance of delineating errors within a structured framework to improve translation quality. Moreover, contrary to the preliminary assumption that iterative processes involve significant computational costs in terms of both financial and temporal resources, our empirical findings demonstrate that smaller models achieve equivalent quality with greater speed and efficiency. The official WMT results confirm this achievement in at least three language pairs: English to Icelandic, Serbian, and Swahili, highlighting the potential of multi-agent AI in Machine Translation. The pertinent code and data are scheduled for public release.

References

Vincent Briva-Iglesias. 2025. Are ai agents the new machine translation frontier? challenges and opportunities of single-and multi-agent systems for multilingual digital communication. *arXiv preprint arXiv:2504.12891*.

Daniel Deutsch, Eleni Briakou, Isaac Caswell, Max Finkelstein, Roni Galor, and 1 others. 2025. WMT24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Ahrii Kim. 2025. RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165, Vienna, Austria. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. GPT-4 Technical Report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. ACL.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. ACL.

Ricardo Rei, José GC De Souza, Daniel Alves, Chrysoula Zerva, Alon Farinha, and Alon Lavie. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 911–918. ACL.

Ricardo Rei, Alon Lavie Farinha, Luisa Coheur, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL.

Stuart Russell and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach*, 3rd edition. Prentice Hall, Upper Saddle River, NJ.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231. AMTA.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Appendix

Example	Icelandic Translation	English Back-translation
Translate	Gagnrýnendur létu SEC-stofnunina hafa það á miðvikudagskvöld.	Critics let the SEC institution have it on Wednesday evening.
Postedit	Gagnrýnendur létu SEC blasta miðvikudagskvöld.	Critics blasted the SEC on Wednesday evening.
Proofread	Gagnrýnendur gagnrýndu SEC að kvöldi miðvikudags.	Critics criticized the SEC on Wednesday evening.

Table 5: An example of Multi-agentMT's output. The modifications are marked in blue. The source segment is **"Critics blasted the SEC on Wednesday night."**

Example	Icelandic Translation	English Back-translation
Translate	Með því að lög um lífskostnað lýkur 31. mars 2024, leitast nýtt samráðsgjörningur við að framlengja stjórn á hækkun leiguverðs á komandi ári. Samráðið lýkur þennan mánudag og leggur til að leiga sé sett eftir lægsta þriggja samanburðarþátta: opinberu markaðsleigu; tillögu leigusalans að nýrri leigu; og „eðlilegri“ hækkun sem unnin er úr nýju skammtakerfi með samanburði við prósentu af markaðsleigu.	With the Cost-of-Living Act expiring on March 31, 2024, a new consultation initiative seeks to extend control over rent increases in the coming year. The consultation ends this Monday and proposes that rent be set according to the lowest of three reference points: the official market rent; the landlord's proposed new rent; and a 'reasonable' increase calculated through a new adjustment system based on a percentage of the market rent.
Postedit	Með því að lög um lífskostnað lýkur 31. mars 2024, leitast nýtt samráðsgagn við að framlengja stjórn á hækkun leiguverðs á komandi ári. Samráðið lýkur þennan mánudag og leggur til að leiga sé sett eftir lægsta þriggja samanburðarþátta: opins markaðsleigu; tillögu leigusalans að nýrri leigu; og „eðlilegri“ hækkun sem unnin er úr nýju skammtakerfi með samanburði við prósentu af markaðsleigu.	With the Cost-of-Living Act set to expire on March 31, 2024, a new consultation document aims to extend control over rent increases in the coming year. The consultation ends this Monday and proposes that rent be determined based on the lowest of three reference points: open market rent; the landlord's proposed new rent; and a 'reasonable' increase calculated using a new adjustment system that compares a percentage of the market rent.
Proofread	Með því að lög um lífskostnað rennur út 31. mars 2024, leitast nýtt samráðsgagn við að framlengja stjórn á því hversu miklar leiguhækkanir má gera á komandi ári.	With the Cost-of-Living Act set to expire on March 31, 2024, a new consultation document seeks to extend control over how much rent can be increased in the coming year.

Table 6: An example of hallucination made by Proofread Agent. The modifications are marked in blue. The source segment is **"With the Cost of Living Act legislation ending on 31 March 2024, a new consultation document seeks to extend controls on the level of rent increases that can be levied in the coming year. The consultation ends this Monday and proposes rents be set by the lowest of three comparators: open market rent; a landlord's proposed new rent; and a "reasonable" increase devised from a new taper system using comparison with a percentage of market rent."**

Listing 1: Prompt of Postedit Agent. The use of reference is optional.

```
{source language} source: ```{source sentence}```
{target language} translation: ```{translation}```
(Optional) {target language} reference: ```{reference}```

Based on the source [and reference] and translation enclosed in
   triple backticks, identify only errors in the translation and
   classify each by category.
Categories: addition, mistranslation, omission, untranslated text,
   grammar, inconsistency, punctuation, word order, terminology, and
   style. You must find at least one issue, even minor, stylistic, or
    subjective.
Rate severity from 1 (minor) to 4 (severe distortion). Never select
   entire sentences or long phrases as an error span. Select only the
    exact word or short phrase where the error occurs. Suggest fixes
   *only* for the erroneous parts -- do not rewrite the full sentence
   .

Format:
{
  "<error span>": {
    "category": "<category>",
    "severity": <1-4>,
    "suggestion": "<fix>"
  },
  ...
}
```

Listing 2: Propmt of Proofread Agent

```
Review the given translation for errors. Find errors and correct them
    first. Then, generate five rephrased translations optimized for
   fluency and adequacy in the {domain} domain. Select the most
   contextually appropriate version based on linguistic fluency in {
   target language}, preservation of source accuracy, and adherence
   to professional translation standards. Output only the final best
   translation. Do not include the other versions, reasoning, or any
   additional text. The output must consist of a single sentence only
   .

{source language} source: ```{source sentence}```
{target language} translation: ```{translation}```
```